Luo et al. The Journal of Chinese Sociology (2019) 6:11
https://doi.org/10.1186/s40711-019-0102-4

The Journal of Chinese Sociology

METHODOLOGY Open Access

Big data research guided by sociological theory: a triadic dialogue among big data analysis, theory, and predictive models

Jar-Der Luo[1*], Jifan Liu[2], Kunhao Yang[2] and Xiaoming Fu[3]

* Correspondence:
jarderluo@126.com
[1]Sociology Department Social Science School, and Public Administration School, Tsinghua University, 100084 Beijing, People's Republic of China
Full list of author information is available at the end of the article

Abstract

Computational social science has integrated social science theories and methodology with big data analysis. It has opened a number of new topics for big data analysis and enabled qualitative and quantitative sociological research to provide the ground truth for testing the results of data mining. At the same time, threads of evidence obtained by data mining can inform the development of theory and thereby guide the construction of predictive models to infer and explain more phenomena. Using the example of the Internet data of China's venture capital industry, this paper shows the triadic dialogue among data mining, sociological theory, and predictive models and forms a methodology of big data analysis guided by sociological theories.

Keywords: Big data analysis, Computational social science, Dynamic network, Circle theory, Embeddedness theory

Introduction

Big data analysis[1] has drawn great attention to computational social science. This paper focuses in particular on computational sociology. However, earlier big data analysis focuses only on the practical and treats collected data as simply the population. These works do not emphasize random sampling or causal inference but mainly focus on descriptive statistics and correlation analysis. This kind of big data analysis that centers on data mining (Mayer-Schönberger and Cukier, 2014) usually only answers the "what" questions, but not the "how" (the mechanism of the process's unfolding) or the "why" (causal relations) questions. Without answers to these queries, predictive models derived from relevant research lack the capability to make causal inferences (Rubin 1974). In the most widely cited example of big data analysis, a grocery store is advised to place beer next to the diaper aisle because its cashier data shows a high correlation between the purchase of diapers and the purchase of beer. However, what should be done is to further ask what kind of people these customers are, what their purchase style is, their psychological state when making the purchasing decision, and so on. Only with a theory to answer these questions can a predictive model determine its range of inference. For example, when does a currently valid prediction lose its validity? Will a prediction that is valid in the USA be valid in China as well? Can we make inferences about other products using such purchasing behaviors?

Predictors and behavioral patterns acquired from big data mining, and the results derived from induction can falsify theories behind different existing predictions, but cannot establish any new theory (Popper 1965). We still need to interpret the results from data mining, dialogue with the existing theories, develop hypotheses, and collect "facts"[2] that related academic communities have accepted in consensus. Then, we test the hypotheses with "facts" and finally obtain approval theory from scholar communities (Lakatos 1980). Only when we use theory to build predictive models can we infer new "facts" at different times, in different settings, regions, and cultures, and thus construct predictive power capable of inference (Galison 1987). Simply put, it is the theory that makes inferences, not data or the results of data mining. Therefore the dialogue between sociology theories and data mining is crucial for predictive models' inference capability.

Developments in computational social science bring social science theories into data mining, especially that of big data. On one hand, big data can test hypotheses derived from theories. For example, measures of social capital using the frequency of telephone calls within a large region show that a community's economic development is affected by its social capital (Eagle et al. 2010). At the same time, when a theory is less than clear, data mining provides clues for its development. We can interpret the outcome from data mining so as to have dialogue with theories that may explain the phenomenon and thus develop a new theory.

On the other hand, theory can guide the direction of data mining. For instance, Dunbar maintains that the social network is divided into multiple circles based on the level of intimacy. He analyzes social network data (Dunbar et al. 2015) and develops a calculation for what we call "Dunbar circles." Moreover, theoretically informed qualitative and quantitative research can be used in data collection to correct the result of data mining, that is, the process of ground truthing. Data obtained by conducting qualitative and quantitative research guided by social science theory and methodology for targeted phenomena of data mining is called "ground truth." Ground truth is originally a term from remote sensing research (Seager 1995) that refers to investigating what exactly is the object (truth) in a satellite image of the ground. When used in data mining this concept means examining whether a predictor or behavior pattern obtained in data mining actually exists in real life, and how much difference there is between the mined phenomenon and the real life "fact." In other words, theoretically informed investigation can provide a ground truth to test the outcome of data mining. For example, Kosinski uses big data from Facebook to calculate five major dimensions of the personalities of Facebook users (Kosinski et al. 2016). He first uses survey methodology to collect real-life personality measures from a group of people as ground truth and then records these people's online behaviors on Facebook for data mining. This empirical ground truth can test the validity of the results of data mining (Kosinski et al. 2016).

Once a theory is "proven"[3], we can use the theory to build predictive models, which can not only improve predictive power with a certain level of accuracy, but can also figure out inferences through theoretical deduction. If either a possible increase of the predictive power or a new inference can be made, this indicates that the theory still has room for improvement. The researcher therefore engages in another round of dialogue between theory, data mining, and predictive models. Enlightened by the new mining, by interpreting the outcome of mining and converse with possible relevant theories, we

can correct the previous theory and retest with new data (big data, survey data, or secondary data). Similarly, the corrected theory can propose new predictive models, deduce new "facts," or, of course, lead to another round of data mining.

Figure 1 illustrates the above process. Simply put, theory contributes to big data analysis by providing ample new topics, pointing to new research directions such as the five dimensions of personalities, the Dunbar circle, and community social capital. At the same time, theoretically informed qualitative and quantitative research can provide the ground truth for the results of data mining and increase the accuracy of the mining outcome. Big data in turn can be used to test a theory, obtain new theories, or correct existing theories through interpreting the mining outcome and engage in dialogue with other theories. Theories then further guide the construction of predictive models, and the latter deduce new "facts." Either in terms of time, cultural environment, or different settings, new "facts" will have corresponding new data. This perpetual spiral continuously corrects theories and extends inferences into a wider field.

In other words, big data and data mining cannot independently make inferences since their results can only make a pragmatic prediction of limited space and time. It is theoretical deduction that helps us make broader inferences and extend the direction of big data analysis in repeated rounds of the triadic dialogue shown in Fig. 1, and thereby continuously improve social science theories, obtain better predictive models, and infer new phenomena in a wider range of fields. In what follows, the authors illustrate this triadic conversation process in its entirety, using the co-investment network (or syndication network) data of a group of venture capital (VC) firms[4] as an example.

The impact of big data on theoretical development

An important step in theoretically guided big data analysis is to turn online, electronically stored data into variables needed to develop theories. Unstructured data is not naturally fit for theoretical deduction or testing. Methods like dynamic network modeling need to be refined, effectively organized, and well structured. To have an effective dialogue between big data and theory, we need a process to structuralize data through algorithms. For example, when analyzing stratification of Chinese society, we need to

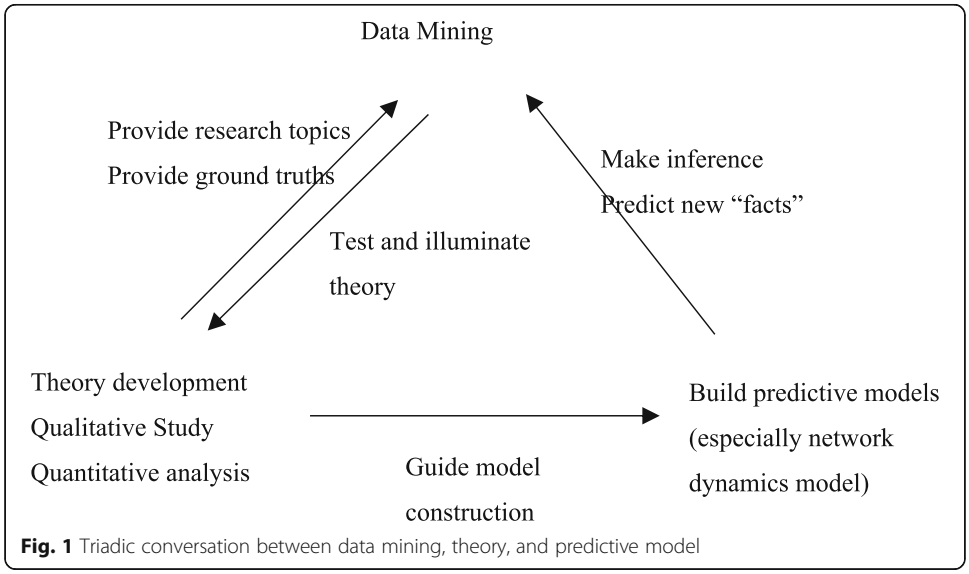

Fig. 1 Triadic conversation between data mining, theory, and predictive model

understand every individual's social economic status (SES). Established variables from previous sociological research include the individual's education, income, wealth, occupation, and reputation. To take this a step further, family background becomes important, and so the career and education of parents could be included. Only then can we derive a complicated SES measure. None of these, however, can be found in online, electronically stored data. Therefore, we may use survey data of SES as ground truth and look for possible behavioral outcomes in the big data, such as residence location or purchases of products. We can then trace this person's moving tracks to find their location at night (possible residence location), their location in the day (possible office), property price in that location, web surfing history (profiling their purchasing style), and online purchasing history (profiling their income). This information can be found in electronic data, and we can thus derive an algorithm method to predict this person's SES. Based on the ground truth obtained from survey, we look for the algorithm to turn big data into structured data, which can thereby be used by theoretical testing and deduction. Currently, the amount of survey data is very limited. When we have an algorithm that can turn big data into theoretically effective measures or, even a step further, deduce a behavioral pattern through the theory, the algorithm and the pattern can be applied to all netizens. The amount of the data becomes enormous.

To summarize, theoretically guided big data analysis has a huge body of data sources, but the amount of data can be big or small after a process of refined and effective organization, which makes the data a good fit for theory development. For example, when we are able to match the SES score of 10,000 surveyed persons to their online behaviors and get an algorism with satisfactory prediction accuracy, we can make inferences regarding all netizens that are similar to these 10,000 individuals. If the 10,000 people are not randomly sampled but belong to a certain social category, inference can only be made to this social category. Millions of online data points could then possibly be turned into structured data good fit for testing and developing theory.

The following sections of the paper take the structure of venture capital firms' co-investment network (or syndication network) as an example of the triadic dialogue. We later explain the content of this case study in detail. Looking only at the part of data collection first, in the current backdrop of high-level digitalization of the economic and financial system, it is actually very easy for researchers to obtain the investment data of venture capital firms. Abundant and detailed information about investment behaviors can be found in numerous financial reports of listed companies, economic news, and publicized material from venture capital firms. However, it is less easy to construct a syndication network from such information. Even after collecting big data with Web-crawling technologies, the data usually remains a collection of sparse investment events. This is the circumstance that the venture capital firms in this paper confront[5]. Matching these events and forming a syndication network is in itself a time-consuming job. Especially for investment events that commonly lack some information, the social science researcher needs treatment of missing values for the validity of the theoretical development.

The examples presented above clearly show that big data does not contain the variables needed for our research (syndication network and corresponding VC companies' indexes) and must thus be handled with a number of methods to structuralize it. This process is far more complicated than typical structured data cleaning. For example,

data matching is a seemingly simple step. In the original data of venture capital firms, a firm's name is typically recorded either in full or in abbreviation in different electronic sources of data and must be recognized and matched. However, there is no certain pattern for such matching. For example, not every firm abbreviates its name as the first two letters of its full name. As a result, even a step as simple as data matching has to make use of multiple techniques, like natural language process and word parsing. In other words, unlike the case of finding individuals' SES, it is not only a job of algorism design in the process of structuralizing data but requires many steps involving labor-intensive work as well. Because of this issue, this study adopts Zero2IPO Research database as the base, from which we collect more online information to clean the missing values and transform the investment event data into venture capitals' syndication network.

In the process of structuralizing data, we first need to determine the missing value in Zero2IPO. For example, for an investment event that lacks information about the time of the investment, researchers need to first conduct an online search of other key information of that time, such as the amount of the investment, the currency used, the place of the transaction, or the information of the recipient, so as to match the corresponding investment events and fill in the missing time data. An algorism must then be designed to match all investment events so as to list the co-investment events and syndication ties. Structuralized big data thus provides us with key theoretical variables that can be used in theory testing and causal inference and helps us use dynamic network methods to build predictive models. The data used in the following paragraphs are essentially this kind of structured data that we have integrated from massive electronically stored data.

In showing the abovementioned triadic dialogue, this paper uses dynamic network modeling as an exemplar of predictive models. The reason is twofold. On the one hand, social science theories have guided the process of data mining and provided more directions for big data analysis. On the other hand, not only can big data be used to test theory and shed light on theory development, but it has also extended the direction of theory construction, especially that of dynamic complex system theory.

The evolution of a dynamic complex social system ought to be a co-evolution of individual behaviors and overall social network structure (Padgett and Powell 2012) . The previous difficulty in repeated collection of wide-range, long-term data can be remedied by unstructured data of electronic footprints. Previous data collection of an ego-centered network, despite the ability to obtain a wide-range random sample, only obtains individual social network conditions, not enough to determine the whole network structure of the wide range. Whole network survey data can be used to analyze the entire structure of a network within a certain range, but this range is compressed to a very small network when using previous research methods. It has been extremely difficult to collect whole network information for several hundreds of people, let along a huge social system that contains millions of people (Wasserman and Faust 1994). Collecting data on network dynamics is even more difficult since people become cautious when asked repeatedly about their personal interactions. It is very difficult to obtain comparative static information for three to five time points (Burt and Burzynska 2017), let alone dynamic network information.

The emergence of social network websites and APPs such as Facebook, Twitter, QQ, and WeChat has completely changed the picture. For more than a decade, the personal networks of billions of people have been recorded. Data on network structure evolution can now be obtained over tens or even hundreds of time points simply by refining and organizing monthly or quarterly electronic footprints data. To this extent, the emergence of big data turns the construction of network dynamics theory and testing its hypotheses from nearly impossible to achievable.

Why is it so important to construct network dynamics theory and build predictive models on it? In both natural and social sciences, complex theory was developed to correct the reductionist tendency of previous theories (Prigogine 1955). The best-known discussion in social sciences is Granovetter's questioning of "under-socialization" and "over-socialization" (Granovetter 1985). The former refers to equalizing collective behavior to the linear summation of individual behaviors, i.e., the collective depends on the individual. The latter refers to subjecting individual behavior entirely to the shaping power of the collective, i.e., the individual depends on the collective. In fact, both presume that individuals are atomized. Such reductionist oversimplification ignores the reality that a collectivity is not the simple sum of individuals but collective behaviors are produced by individuals' binding together and forming large-scale complex social networks. It is the aggregate effect of individual behavior and social network structure that produces collective actions (Granovetter 2017).

Coleman (1990) expresses similar arguments. As shown in Fig. 2, the reductionist view explains collective outcomes with collective elements (Process 4), and individual outcomes with individual elements (Process 2). Process 1, which explains individual outcomes with collective elements, is over-socializing, while Process 3, which explains collective outcomes with individual elements, is under-socializing. Coleman points out that such explanations overlook interpersonal interactions, relations, social networks, and the structure of networks. From the social-network point of view, Process 1 consists of four types of research (Luo et al. 2008), and collective powers can be conceptualized as field forces, including informational and normative field forces (DiMaggio and Powell, 1982). The first line of research argues that collective powers affect individual relations and the formation of personal networks. The second maintains that these relations and egocentric networks influence individual behavioral outcomes through either the interactive effect among friends or social capital that the ego gains from the

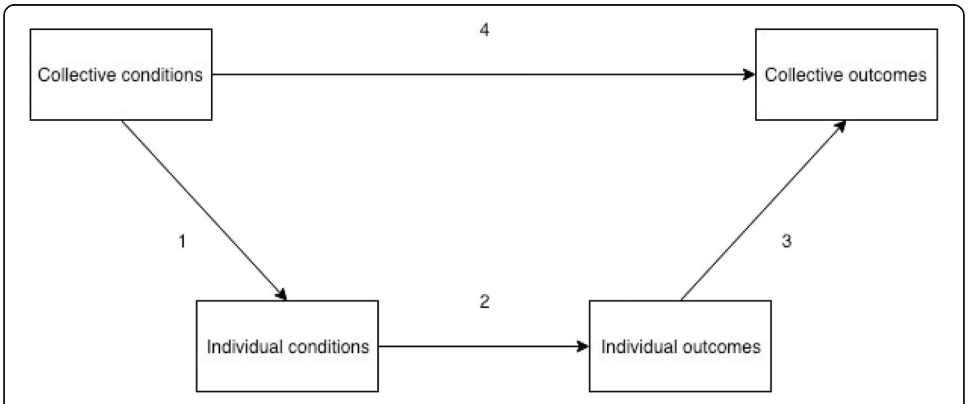

Fig. 2 Illustration of Coleman's argument

network (Lin 2001). The third type of research thinks that field forces affect the changes in broader networks surrounding the individual and thereby changes their structural position in the network. The fourth type argues that individual structural positions, such as structural holes (Burt 1992) or centrality in a closed network (Brass and Burkhardt 1993), also influence individual behavioral outcomes.

Process 3 also includes three different lines of research (Luo et al. 2008). The first addresses the change in network structure induced by individuals' cutting or building relations (Powell et al. 2005), e.g., a topic of network dynamics. The second explores how collective actions emerge from the evolution of network structure and human actions (Padgget and Powell 2012). The third argues that long-term, continuous, large-scale, and influential collective actions will eventually form new field forces and become the powers that shape individual relationships and structural positions in Process 1 (DiMaggio and Powell 1982).

To summarize, we can clearly see that Process 3 is where big data can make the greatest contribution. Research on network dynamics and the emergence of collective action from the co-evolution of structure and action has been filled with theoretical speculations while information to test the theory has been scarce. It has thus been difficult to develop and improve the theory in more depth and details. The use of big data has completed the research loop in Fig. 2, enabled the analysis of nonlinear developments like dynamic changes of large-scale complex networks; the emergence of crucial collective actions such as important innovations, social movements, and revolution breakouts; extranormal evolutions of complex social systems such as financial crises, sudden change of business cycles, and social transformations; and the transition of economic, social, and political institutions. Simply put, as argued above, social science theories can guide the development of big data analysis, discover new topics, collect ground truth, supervise the result of data mining, and conduct broader inferences. The addition of big data also extends the room for development for theories, making it possible to explain and test topics that have been difficult, producing a new frontier for theoretical development.

Why and when does the business cycle of an industry take a sharp turn? When does the turning point come? Is there any pattern or indicator? These are topics of interest that complex theory strives to answer. However, before answering these questions, we need to ask not only how the behaviors of industrial actors change, but also what the industrial network structure looks like and how it evolves. We take the venture capital (VC) industry as an example to consider the structural analysis of the VC syndication network.

Theory-informed ground truths

When trying to understand the structure of syndication networks that VC firms form, big data analysts often employ community detection, which divides the entire industrial network into several to a dozen communities. What is the meaning of these communities? What type of investor does each community represent? What is their investing behavior? Why do they cluster together? For example, when thousands of firms in the VC industrial network are divided into ten communities, we must ask whether the real world can be divided similarly, what meaning each community represents, why a particular group of VC firms cluster, and how they might

evolve. For these questions, *guanxi* circle theory provides a direction of research in the Chinese world (Luo 2012; 2016).

In Chinese daily life, guanxi circles, or "small circles," are normally defined as egocentric personal networks at work—"strong ties developed from egocentric social networks that typically consist of a central node (or a very small group of central nodes) and their pseudo family and familiar ties" (Luo 2012)[6]. In the VC industry, a small circle refers to a syndication network formed among a VC firm and those who frequently cooperate with it. Moreover, each investment has a leading investor who is in charge of initiating an investment plan and holds a seat on the investee's board of directors, while other investors are followers. In the first few rounds of an investment case followers can usually enter the co-investment only at the leading investor's invitation. As such, an investor who often leads investments draws a group of followers with differing degrees of closeness; those closest become the small circle centered around the leading investor (Luo 2014).

Although every investor would like to build their own circle and many of them have large or small circles, only those large VCs that frequently invest or possess industrial-specific knowledge often lead investments, and only their circles are influential in the VC industry. We therefore call them "industrial leaders." The key to our understanding of the whole industrial network is to find the industrial leaders as well as their circle members and observe the cooperation and competitions between them in which their circles evolve (Gu et al. 2019). Therefore finding an algorithm to calculate "industrial leaders" and determining their mechanisms for inviting syndication partners are crucial entry points. Identifying an industrial leader is equivalent to finding an important node in a network. Is its degree centrality, its betweenness centrality, or its investment amount more important? We can get a more-precise algorithm if we take ground truth as a target and look for an algorithm closest to it.

We used the Delphi method, a qualitative methodology of interviewing experts, to collect ground truth. Our research team first calculated the k-shell value[7] of each VC firm, which ranges from one to fourteen, ranked all values, and handed them to four industrywide recognized experts, most of which are CEOs of industrial leading firms, for assessment of industrial leaders. Forty-two VC firms that received consensus approval were chosen as ground truths. Some with relatively higher k-shell values did not make the cut. This speaks to the possibly considerable deviation if we tried to find out these leaders with only one or two indexes, which in the experts' opinion would have been very different from the facts in the real world.

Using these forty-two leading VC firms as ground truths, our algorithm for data mining involves five network indexes—degree centrality, the H index, the k-shell value, eigenvector centrality, and local rank, as well as two non-network indexes—the number of investing popular industries and the number of investing industries. Conducting clustering analysis with these seven indexes, we ended up with finding 35 out of the 42 industrial leaders. The predictive accuracy rate was 0.965, and the recall rate 0.83 (Luo et al. 2018), both higher than outcomes of mining without ground truths, regardless of indexes used.

After studying the industrial leaders that we have mined out, we can go back to the abovementioned example questions. For instance, when thousands of firms in the VC network are divided into ten communities, we can interpret their structure and

understand each one's characteristics: Which communities are the circle of a certain leader? Which are circles formed by several leaders bonding together? Which are relatively bigger communities that have a small-world network structure (Watts 1999) and consist of multiple circles? From the characteristics of these industrial leaders, we can also understand whether a community is composed mainly of foreign VC companies, state-owned capital, or domestic private VCs. We can also determine whether their investment targets are monotonic or diverse, or which industries are included. These questions and others help us further understand the structure of the industrial network.

Data mining and theory development

In addition to finding the leaders, what are the mechanisms by which they invite syndication partners? Data mining has a role in answering the question about the origin of bilateral co-investment relationships in the VC industrial network. In the database mentioned above, fully eighty-one indexes can be derived only for network relations, but which of them are more important? Researchers computed the seven best predictors using the structural balance-based factor graph (SBFG)[8] model—same nationality, number of shared neighbors, betweenness centrality, relationship distance, same type of ownership, number of investment fields, and number of common investment fields (Zhou et al. 2016).

Analyzing these mining results, we clearly see two types of the most influential factors: network structure variables (ranked 2, 3, and 4, respectively) and similarity variables (ranked 1, 5, and 7, respectively). This discovery is quite different from research findings about American biotech VC firms in which the importance of similarity is mainly controlled by other variables and the result is thus insignificant (Powell et al. 2005). This reveals the difference between the syndication behaviors of the Chinese VC industry and those of American ones. Based on these results of data mining, we introduce the embeddedness theory to explain the emergence of this syndication behavior[9]. This is the process of dialogue between existing theories and data mining outcomes.

The embeddedness theory includes relational embeddedness and structural embeddedness (Granovetter 2017). Gulati (1999) brings the relational embeddedness theory into his research on strategic alliances and argues that the more two strategic partners cooperate the stronger the trust they cultivate and the more they are in harmony when cooperating. It is thus more likely that they will cooperate again (Gulati 1999). Based on this, we propose the following hypothesis:

Hypothesis 1: The experience of cooperation between two venture capital companies and the likelihood of their cooperating again are positively correlated.

This is a reasonable hypothesis for the circle effect in the Chinese VC industry. The centered VC of a circle needs a group of particularly close partners—its "team." In the process of starting an investment plan, a strong and capable team is often required to act quickly on the emergence of opportunities, mobilizing potential resources and grasping the chance (Granovetter 1995; Burt and Burzynska 2017). At the same time, a tight group has lower moral risks when facing a highly uncertain environment because of the supervisory effect of an internal dense network, or the trust of a third party can be used as a trustworthy promise to reduce transaction cost in the cooperation

(Granovetter 1985). As such, frequent partnership results in more-frequent cooperation, and the firm becomes a core member for the circle leader.

Structural embeddedness refers to how an actor's behavior and its result are affected by their position in the social network. In discussing the cooperative relation between two actors, we need to take into account their relative structural positions. Relationship distance affects the actors' cooperation in two ways. First, trust can transmit (Burt and Knez 1995). Two adjacent nodes—actors who have past experience of cooperation—are likely to cooperate again because they know each other very well. Between two nodes with a distance of two steps—i.e., friend of my friend—trust is born from the endorsement of the friend, which makes cooperation easier. Second, birds of a feather flock together, meaning that friends of my friends are highly likely to be similar to me. As they meet in various social situations, their originally indirect relationship is likely to become a direct one and produce greater chances of cooperation (Granovetter 1973). The farther the distance and the less transmission effect of trust, the lower the likelihood is of two nodes becoming direct friends. Based on this, we propose the second hypothesis:

Hypothesis 2: Relationship distance and future co-investment relations are negatively correlated.

The number of shared friends is an important factor of data mining because the transmission effect of trust fades away quickly and dies out after three steps—in other words, endorsements of my friend's friend are no longer trustworthy. At the same time, nodes with a distance of more than three steps away from each other are unlikely to gather and get to know each other directly, and therefore the chance for cooperation decreases to zero. The more shared friends, the stronger the transmission effect of trust, and the higher the chance to meet the higher the possibility of cooperation is. We therefore propose the following hypothesis:

Hypothesis 3: The more common neighbors that two VCs have, the higher the possibility is of their cooperation.

Putting Hypotheses 2 and 3 in the guanxi circle phenomenon of the Chinese VC industrial network, an environment with extremely high uncertainty, we realize that the leader of a circle needs not only close members but also a large number of weak ties to access more opportunities. The small circle should alter from time to time between strongly coupling and weakly coupling networks to bring about different resources (Granovetter 2002). By having more partners with which it has cooperated, a VC firm grasps more investment opportunities since friends of friends bring in different resources. As such, we see the leader of a circle build multilayered networks consisting of both close core members and periphery insiders to have both mobilizing ability and remain open to more opportunities (Luo et al. 2017). Based on some invitation mechanisms, a leading VC selects various layers of partners to form its circle.

Using frequency of cooperation, relationship distance, and the number of common neighbors between two VC firms as independent variables, researchers tested the hypotheses stated above after controlling for variables based on previous theories such as accumulated advantage and similarity of investment areas. The results support all three of the above hypotheses (Luo et al., 2018a, b, c).

The example above reveals that the result of data mining can shed light on the construction of theory, but the process of theory development still comes from dialogue

with other theories, logical deduction, proposal of hypotheses, and testing the hypotheses against the data.

Theory-informed dynamic models

The following presents a case of dialogue between confirmed theory and predictive model building[10]. Simply by constructing predictive models based on the causal model deducted from guanxi circle theory and embeddedness theory as described above, we can make inferences about which two companies are likely to form a cooperative relationship at different time points or in different industries, or in a similar cultural environment. However, this is not enough for network dynamics. The driving effect of introducing big data into theory development is reflected chiefly in studying complex dynamic systems. To dive into this research area predictive models of network evolution must be built.

Without direction from theories, network dynamic models often control basic network statistics such as network scale, rate of growth, and network density and build random graph models as basic models by letting nodes randomly form lines connected to other nodes. Other network statistics of interest, such as the number of closed triads and other types of motifs, are then added into the model as independent variables to see by how much the accuracy of predicting future network structure has increased compared to the basic models.

Since there are investors and investees (Gu et al. 2019) in the VC industry, the researchers built a two-mode random graph to model networks evolution in fourteen time stamps in comparison with the real structures in the 14 years between 2000 and 2013. The starting year, 2000, had seventy-five investors and 375 investee firms; both groups increased their number by 30% annually. Investors were further divided into nine categories and fitted to a three-by-three table: {high, medium, low} investment frequency × {high, medium, low} syndication tendency. According to real network statistics, the investor with the highest frequency made 5.047 investments each year (five times each time stamp in the model), the medium-frequency investors made 0.796 (four times every five time stamps), and the low-frequency investors made only 0.26 (once every four time stamps). With these control variables, investors randomly chose investees in every period based on their investment frequency. The model gets a co-investment when two investors invested in the same investee. The accumulation of these co-investments produced an industrial syndication network, which is our Model 1, stochastic investment model, or the baseline model.

When circle theory and embeddedness theory were brought in, investments were no longer random. Instead, the leading investor invested first, then partners were invited to participate in the co-investment in the patterns proposed in Hypotheses 1, 2, and 3. Moreover, investors were also divided by their tendency to syndicate—i.e., the number of an investor's co-investments divided by the number of his/her total investments. Those with a high tendency chose to invite other investors 90% of the time, while medium-tendency investors had cooperation three out of five times, and low-tendency investors only invited followers once every five investments. Each stimulation consisted of two rounds; in the first, investors randomly invested in investees, and in the second, the leading investors invited others to join in, following the above rules. Under the

relational embeddedness theory of Hypothesis 1, the higher the frequency of past cooperation, the higher the likelihood of future cooperation. This is our Model 2.

Our Model 3 takes into account the structural embeddedness proposed in Hypotheses 2 and 3. We set the probability of cooperation to 0 for VCs more than three steps away from each other. The probability is lower for those connected by two steps than it is for those with direct ties. The probability functions for VCs with direct ties were set based on Hypothesis 1. In addition, the more common neighbors two VCs have, the higher the likelihood of their future cooperation. After running fourteen periods of stimulation models and comparing them to the accumulated real network, the following results were obtained. Apparently among macro-level network indexes, as shown in Fig. 3, the degree distribution of the industrial network is better in Models 2 (relational embeddedness model in the figure) and 3 (structural embeddedness model in the figure) than in Model 1 (stochastic investment model in the figure). In other words, models that include the relational embeddedness model (Model 2) or both relational and structural embeddedness (Model 3) are considerably closer to the real network than in the random model with only controls. Also, Model 3 is superior to Model 2 in terms of fitness.

Comparing the micro-level network statistics of all motifs, the baseline model has very poor predictive power whereas the prediction of network structure is vastly improved in Models 2 and 3. In other words, models built under the guidance of theory have more-accurate predictions than does the random model with only controls. Also, Model 3 is superior to Model 2 in terms of accuracy (Gu et al. 2019, Table 1).

Simply put, the prediction of network dynamic evolution using only network statistics as control variables is less than satisfactory, but the addition of theories greatly

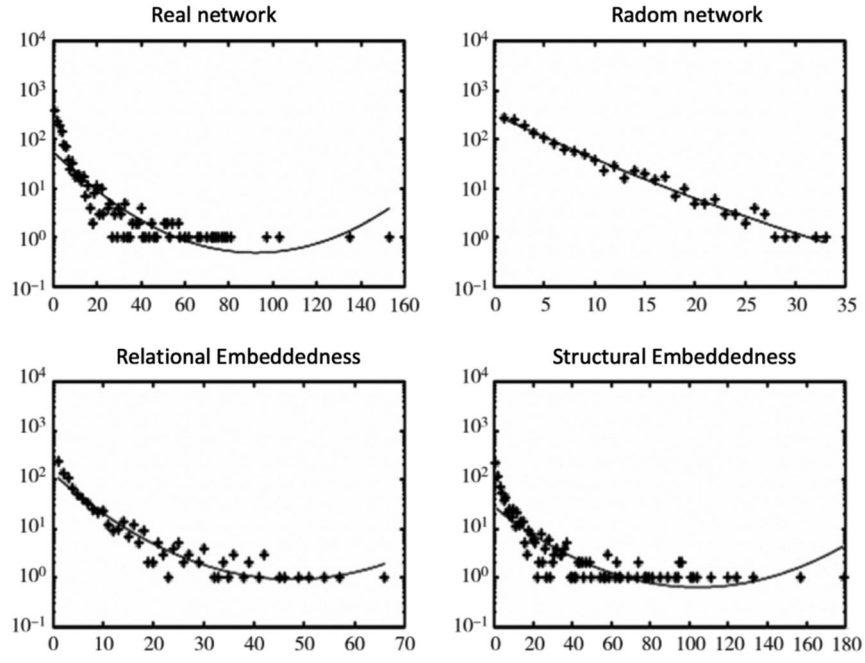

Fig. 3 Macro level network index—a comparison of degree distribution

Table 1 The comparison of various motifs among predictive models

Comparison between micro level network indexes				
Pattern of motif	Real network	Random network	Relational embeddedness	Structural embeddedness
	7668	634	6646	7566
	633815	14642	441325	469410
	0	551	0	32126
	58765	127	51026	53060
	10258	50	5041	5359

Source: Gu et al., 2017.

improves predictive models' power. The above example shows that introducing guanxi circle theory and embeddedness theory into the model produces Model 2 and Model 3, which are much more accurate in predicting the evolution of the VC industrial network than the baseline model, which only uses some network statistics as controls.

Going back to the study of complex system dynamics and nonlinear evolution, what is the meaning of this network dynamic prediction?

As shown in Fig. 4, the guanxi circle phenomenon makes the Chinese VC syndication network emerge as a special type of structure. Industrial leaders build their circles, which makes the whole industrial network full of "caves." There are many long-

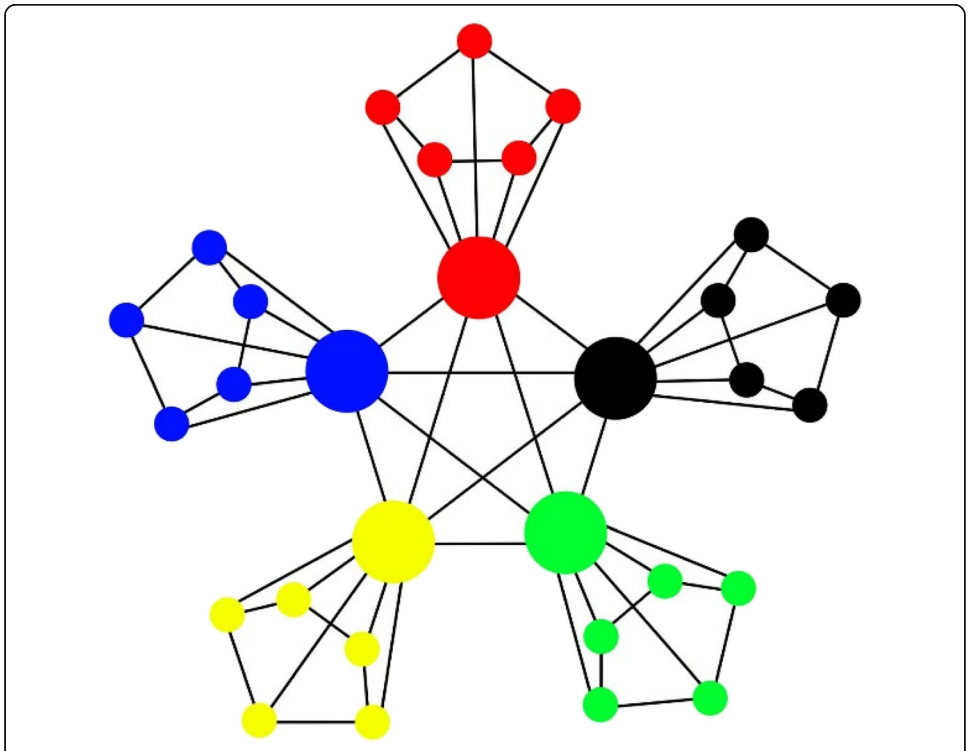

Fig. 4 An example illustrating the small-world network. A small-world network with an elite-clique. Elites are not only the centers of caves but also bridges among caves

distance bridges connecting these circles, which make the structure a "small-world network" (Watts 1999). Additionally, these long-distance bridges bond together to form an "elite clique" (Useem 1984), so the structure turns into a small-world network with a centered elite clique. In the Chinese VC industry, circle members are often not encouraged to play the role of bridges; those circle leaders connect to each other, which forms the structure shown in Fig. 4. The real network statistics provide evidence of the structure of the Chinese VC industry, and Model 3 effectively shows the emergence of this structure in the process of network evolution (Gu et al. 2019).

Can we use the network dynamic model as an evolution mechanism to predict some nonlinear system transformations such as financial crises? It still has a long way to go before it can explain phenomena emerging from the evolutionary process of a complex system. Emerging phenomena such as financial crises are the result of the interactive influences between network structure and behavior, but we have yet to learn about the actions of related actors, as well as the motivations and evolutions of such actions. Therefore, we lack understanding about the causal mechanism of this systematic nonlinear evolution. In addition, predictive models derived from theory must be tested for validity and reliability with more "facts." What is presented here is only one example, but the inferential capability of this predictive model can only be obtained through collecting information on more emerging network structures and putting the model into the test. We have just begun the long journey to study Process 3 in Fig. 2, the process by which individual behaviors are integrated into macro-level collective behaviors. This in turn pushes a social system forward in the interactions between behaviors and network structures. Nevertheless, building predictive models of network dynamics is a firm step in the right direction.

Conclusion: the triadic dialogue between data mining, theory, and the predictive model

The goal of this paper is to illustrate the research methodology of computational social science, the key to which is the triadic dialogue between data mining, theory development, and predictive model building, as shown in Fig. 1. For big data research data mining must dialogue with social scientific theories and build a predictive model to predict new "facts" so as to complete a big data research capable of making inferences. Simply mining data is only the start of the entire cycle of research. Short-term, practical predictions can be derived from it, but this is not enough for theory development. However, inferences must be made from theoretical deduction.

This paper analyzes the VC industry as an example to illustrate the triadic dialogue shown in Fig. 1. The "Theory-informed ground truths" section illustrates the process in which guanxi circle theory sets discovering an algorithm for industry leaders as the topic for data mining in the VC industrial network, collects ground truths through qualitative methods guided by theory, supervises the search for algorithms, and corrects the algorithms' accuracy.

The next section, "Data mining and theory development," explores the reverse situation—i.e., the contribution of data mining to theory. Data can be used to test theory, while interpretation of mining outcomes can dialogue with other theories to illuminate the possible direction for the development of new theories.

The third section, "Theory-informed dynamic models," discusses how to build predictive models with the guidance of theory and reproduce real-world network structures with predictive models. This predictive model illustrates the emergence of a special type of network structure in the Chinese VC industry shown in Fig. 4.

There are certainly flaws in the process of constructing theories and predictive models. For example, our algorithm to find industry leaders has room for improvement. In the current ground truths, seven industrial leaders have not been found. Finding them and the reason they are missing, and correcting the algorithm are some of the next steps. Similarly, the predictive power of the network dynamic model that contains guanxi circle theory and embeddedness theory can also be enhanced. Theories continue to develop, and models continue to be corrected.

Other than further developing guanxi circle theory and embeddedness theory in the study of industrial networks, dialogues with alternative theories can also be a fruitful start. For instance, the "Data mining and theory development" section highlights the importance of similarity to the Chinese VC industry when we mine data on the origin of syndication networks. Theories and predictive models should be developed to see which theory, network or similarity theory, has the better predictive power. Of course, we cannot rule out the possibility that the combination of alternative theories produces the best predictive model, and therefore an integrated theory is constructed. Other research topics and new "facts" from inferences are worth deeper discussion. A new round of triadic dialogue often brings about unexpected development of theories and predictive models.

To summarize, all the aforementioned reasons set in motion another round of triadic dialogue, as shown in Fig. 1—finding new topics, collecting more ground truths, mining more data, interpreting new results, carrying out theoretical dialogues, developing hypotheses, testing theories, using newly improved theories to guide the construction of new predictive models, and using new material to test new predictions, and so forth. Round after round of triadic dialogue continuously improves theories and raises the accuracy of predictive models and the range of their inferences.

We should note, however, that the theoretical development we have alluded to in the Process 3 of Fig. 2 is only the start of a long road of discovery. This paper simply illustrates a research method while touching on the construction of network dynamic models. If we want to explain and predict the nonlinear evolution of a complex system; "emerging" phenomena like financial crises; major innovations; social movements; new thoughts; institutional changes; the transition of political, economic, or social systems; or even revolutions, research like this is far from sufficient. As argued, we need to develop theories to explain the actions of related actors as well as the motivations and evolutions of such actions. More importantly, we also need to investigate the interactive effect between network structures and the dynamic evolution of behaviors as well as how they act together to influence the whole system. Despite the long road and significant tasks, the addition of big data makes available the collection of relevant material and opens a new research area for theory development.

Earlier big data analysis only focused on short-term and practical purposes. Its analytical strategy was to treat collected data simply as the population and mainly focus on descriptive statistics and correlation analysis. This kind of big data analysis focusing on data mining obtains predictors and behavior patterns that have only limited inferential

capability. Only when we answer the "why" and "how" questions can we make broader and more-precise inferences. Now that computational social science, particularly computational sociology in this paper, has brought social science theories into big data analysis, the two can implement each other and make progress simultaneously. In addition, theories bring more topics to big data analysis, and theory-informed qualitative and quantitative research provide ground truths for correcting data mining. On the other hand, big data provides material to test theories, and the outcomes of data mining can shed light on the development of new theories. Theories that pass testing then can inform the construction of predictive models, point to the boundary of the models' inferences, and thereby predict more new "facts" within the boundary of the model.

In future research, the introduction of big data may extend the boundary of social scientific theories into Process 3 in Fig. 2—how individual behaviors are integrated into collective behaviors, transform to macro-level field forces, and especially how emerging phenomena are produced in a complex social system—and finally bring about nonlinear transformation of the system. Theorists have long been able to only speculate on but not test their theories due to lack of data. They have therefore found it difficult to develop and improve theories in more depth and detail, let alone predict evolutionary phenomena with models. Big data makes it possible for this kind of complex dynamic model to extend theoretical boundaries. The triadic dialogue integrates computer scientists' advantage in big data mining and social scientists' advantage in theory and qualitative and quantitative research. More researchers are needed who can construct complex dynamic models. Interdisciplinary integration, especially those who exhibit skill in both social sciences and natural sciences and can mediate conversations, are key to big data research. The biggest challenge to the research community will be taking down disciplinary walls and learning from each other with an open mind.

Since a large amount of human actions now occur on the Internet, online behavior and real life are linked in a deducible manner. The big data that records everyday electronic footprints is already there, and our challenge is to correctly and best use the data in a computational social scientific way.

Endnotes

[1]In this paper, the term "big data" refers to the opposite of structured data (databases of variables that are built by conducting social surveys and/or organizing secondary sources). It refers to the unstructured data of electronic footprints that an actor leaves after online behavior.

[2]We put the term "facts" in quotation marks to avoid debate about the existence of objective fact. Testing of alternative theories is based on data that is inter-subjectively acknowledged in relevant scholar communities.

[3]We also put "proven" in quotations marks to clarify that the theory has the highest explanatory power among relevant competing theories when compared to acknowledged "fact" in the academic community. Competition between theories is based on commonly acknowledged "facts," avoiding the theoretical empirics pointed out by logical empiricism. When the hypotheses from a theory are supported at a certain significance level the theory is then supported. Refer also to footnote 2 and Hempel (1966).

[4]The original source of this data is investment information published by venture capital firms that we collected on the Internet. If two VCs (with the exception of private

equities and angel investments) announce investments in the same company at the same time point, we count it as a co-investment. The original data was primarily cleaned by Zero2IPO Research (PEdata database), and our research team supplemented it with necessary information collected online to obtain the final syndication network data.

[5]That is, the PEdata dataset maintained by Zero2IPO Research.

[6]Yang (1993) divides the Chinese pattern of difference sequence networks into three tiers—family, acquaintances, and strangers. Normally the first two are composed of strong ties while the last are weak ties. Please refer to Yang (1993) for further information.

[7]The k-shell is one of the measures of the importance of a node in a network. It is calculated as follows. First, eliminate every node that only has one tie with other nodes in the entire network; the eliminated nodes have a k-value of 1. Then eliminate each node that has two ties with others; the eliminated nodes have a k-value of 2, and so on and so forth until all nodes in the network are eliminated. The k-value of a node indicates the order of the round in which it is eliminated.

[8]The paper compares multiple algorithms. SBFG ends up the best in terms of both the accuracy of predicting co-investment and the rate of convergence. For the calculations, see Zhou et al. (2016).

[9]This part is adopted from Luo et al. (2014); Luo et al. (2018a); and Luo et al. (2018c)

.

[10]This part is adopted from Gu, Luo, and Liu (2019).

Acknowledgements
Examples presented in this paper come from multiple other papers that the authors have collaborated with others (Luo et al., 2014; Zhou et al., 2016; Luo et al., 2018a, b, c; Gu et al. 2019). We hereby thank all these authors. Moreover, the authors thank the student exchange program between the School of Social Science of Tsinghua University and University of Göttingen in Germany (IDS – SSP – 2017001), which is supported by the Institution of Data Science of Tsinghua University. This article also received support from Tacent Social Research Center, project "Analyzing Ego-Centered Network by Mining of Wechat and QQ Data," Project number: 20162001703.

Authors' contributions
LJD is the leader of all relevant research projects and the main author of this paper. In addition, he is also the first author or co-author of all papers containing examples presented in this article. LJ builds the network dynamics model presented in this paper. YK helps writing the part of big data collection and do some analysis of VC network. FX provides instructions on the methodology thinking in this paper. All authors read and approved the final manuscript.

Funding
The authors thank the student exchange program between the School of Social Science of Tsinghua University and University of Göttingen in Germany (IDS–SSP–2017001), which is supported by the Institution of Data Science of Tsinghua University. This article also gets the support of Tacent Co. Tacent Social Research Center, project "Analyzing Ego-Centered Network by Mining of Wechat and QQ Data", Project number: 20162001703.

Availability of data and materials
The venture capital's syndication network data was transformed from an open database. For the quantitative study in this research, data was mainly collected from the Zero2IPO Research database, which collects a variety of data from the internet and modifies it into a structured form. This dataset provides more detailed and rich information on the Chinese VC industry than the other databases. Investment data includes investment time, investees' name, industry, location and stages in their growth, etc. This data forms a 2-mode network, i.e., a network composed of the arcs from investors to investees. Information from the internet and government documents was used to modify the data set, which aided in distinguishing state-owned VC firms from Chinese private VC firms.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Sociology Department Social Science School, and Public Administration School, Tsinghua University, 100084 Beijing, People's Republic of China. [2]Sociology Department, Tsinghua University, Beijing, People's Republic of China. [3]School of Mathematics and Computer Science, University of Göttingen, Göttingen, Germany.

Received: 28 March 2019 Accepted: 13 June 2019
Published online: 05 July 2019

References

Brass, D.J., and M.E. Burkhardt. 1993. Potential power and power use: An investigation of structure and behavior. *Academy of Management Journal* 36: 441–470.
Burt, R. 1992. *Structural Holes: The Social Structure of Competition*. Cambridge: Harvard University Press.
Burt, Ronald S., and Katarzyna Burzynska. 2017. Chinese entrepreneurs, social networks, and guanxi. *Management and Organization Review* 13 (3): 1–40.
Burt, Ronald S., and M. Knez. 1995. Kinds of third-party effects on trust. *Rationality and Society* 7: 255–292.
Coleman, James. 1990. *Foundations of Social Theory*. Cambridge: The Belknap Press.
DiMaggio, Paul J., and Walter W. Powell. 1982. *The Iron Cage Revisited: Conformity and Diversity in Organizational Fields*. New Haven: Yale University Press.
Dunbar, R.I.M., V. Arnaboldi, and M. Conti. 2015. The structure of online social networks mirrors those in the offline world. *Social Networks* 43: 39–47.
Eagle, Nathan, Michael Macy, and Rob Claxton. 2010. Network diversity and economic development. *Science* 328: 1029. https://doi.org/10.1126/science.1186605.
Galison, Peter. 1987. *How Experiments End*. Chicago: University of Chicago Press.
Granovetter, Mark. 1973. The strength of weak ties. *American Journal of Sociology* 78: 1360–1380.
Granovetter, Mark. 1985. Economic Action and Social Structure: The Problem of Embeddedness. *American Journal of Sociology* 91: 481–510.
Granovetter, Mark. 1995. The Economic Sociology of Firms and Entrepreneurs. In *The Economic Sociology of Immigration: Essays in Networks, Ethnicity and Entrepreneurship*, ed. Alejandro Portes, 128–165. New York: Russell Sage Foundation.
Granovetter, Mark. 2002. A Theoretical Agenda for Economic Sociology. In *The New Economic Sociology: Development in an Emerging Field*, ed. R.C. Mauro, F. Guillen, P. England, and M. Meyer, 35–59. New York: Russell Sage Foundation.
Granovetter, Mark. 2017. *Society and Economy—Framework and Principles*. Cambridge: Harvard University Press.
Gu, WeiWei (joint first author), Jar-Der Luo (joint first and corresponding author), and JiFan Liu. "Exploring small-world network with an elite-clique: bringing embeddedness theory into the dynamic evolution of a venture capital network." *Social Network* 57: 70-81 (2019), https://www.sciencedirect.com/science/article/pii/S0378873318302272?dgcid=author.
Gulati, R. 1999. Network location and learning: The influence of network resources and firm capabilities on alliance formation. *Strategic Management Journal* 20 (5): 397–420.
Hempel, Carl G. 1966. *Philosophy of Natural Science*. Upper Saddle River: Prentice Hall.
Kosinski, M.W., L.H. Yilun, and J. Leskovec. 2016. Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods* 21 (4): 493–506.
Lakatos, Imre. 1980. *The Methodology of Scientific Research Programmes: Volume 1: Philosophical Papers*. Cambridge: Cambridge University Press.
Lin, Nan. 2001. *Social Capital: A Theory of Social Structure and Action*. New York: Cambridge University Press.
Luo, Jar-Der. *Guanxi* and circle—circle phenomenon in the chinese work field. (in Chinese). *Chinese Journal of Management* 2 (2012):1-8.
Luo, Jar-Der. 2016. Guanxi circle phenomenon in the chinese venture capital industry. In *Social Capital and Entrepreneurship in Greater China*, ed. Jenn hwan Wang, 56–71. New York: Routledge.
Luo, Jar-Der, L. Cao, and R. Guo. How does embeddedness influence VC co-investments. (in Chinese). *Jiangsu Social Sciences* 4 (2018): 85-96.
Luo, Jar-Der, Y. Fan, Q. Guo, J. Zhou, J. Liu, and R. Li. 2018. *How to find industrial leaders in venture capital—ground truth in computational social science*" (in Chinese). *Exploration and Free Views* 7, 94–102.
Luo, Jar-Der, Rong Ke, Kuan-Hao Yang, Rong Guo, and Ya-Qi Zou. 2018. Syndication through social embeddedness: A comparison of foreign, private and state-owned venture capital (VC) firms in China. *Asia Pacific Journal of Management*. https://doi.org/10.1007/s10490-017-9561-9.
Luo, Jar-Der, Ray-Chi Li, Fang-Da Fan, and Jie Tang. 2017. "Mining data for analyzing *guanxi* circle formation in chinese venture capitals' co-investment." In *Interdisciplinary Social Network Analysis,* eds. Xiaoming Fu, Jar-Der Luo, and Boos Margret, *177-196.* New York: Taylor and Francis Group.
Luo, Jar-Der, L. Qin, and L. Zhou. Circle phenomenon in the chinese venture capital industry. (in Chinese). *Chinese Journal of Management* 4 (2014): 469-477.
Luo, Jar-Der, J. Wang, J. Zhang, and Z. Xie. The architecture of social network analysis—take organizational theory and management research as examples. (in Chinese). *Society* 4 (2008): 15-38.
Mayer-Schönberger, Viktor, and Kenneth Cukier. 2014. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. London: John Murray Inc.
Padgett, John F., and W.W. Powell. 2012. *The Emergence of Organizations and Markets*. Princeton: Princeton University Press.
Popper, Karl. 1965. *The Logic of Scientific Discovery*. New York: Harper Torch Book.
Powell, W.D., D.R. White, K.W. Koput, and J. Owen-Smith. 2005. Network dynamics and field evolution: The growth of inter-organizational collaboration in the life sciences. *American Journal of Sociology* 110: 1132–1205.
Prigogine, I. 1955. *Thermodynamics of Irreversible Process*. New York: Ryerson Press.
Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66 (5): 688–701.
Seager, W. 1995. Ground truth and virtual reality: Hacking vs. van Fraassen. *Philosophy of Science* 62: 459–478.
Useem, M. 1984. *The Inner Circle*. New York: Oxford University Press.
Wasserman, S., and K. Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.
Watts, Duncan. 1999. Dynamics and the small-world phenomenon. *American Journal of Sociology* 105 (2): 493–527.

Yang, Guoshu. 1993. *Social Orientation of Chinese People—A Social Interactional View (in Chinese)*. Taipei: Lauréat Publications.
Zhou, Yun, Zhiyuan Wang, Jie Tang, and Jar-der Luo. 2016. The prediction of venture capital co-investment based on structural balance theory. *Transactions on Knowledge and Data Engineering* 28 (2): 537–550.

Publisher's Note